ELSEVIER

Journal of Computational and Applied Mathematics 82 (1997) 389–405

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Error estimates using the cell discretization method for steady-state convection–diffusion equations

Howard Swann*

San José State University, San José, CA 95192-0103, USA

Received 25 July 1996; received in revised form 15 February 1997

Abstract

The cell discretization algorithm is applied to generate approximate solutions for some second-order non-self-adjoint elliptic equations. General convergence for homogeneous problems is shown by obtaining suitable error estimates. The method is applied using polynomial bases; this provides a nonconforming extension of the finite element method that can also produce the continuous approximations of an h–p finite element method. Numerical tests on convection–diffusion problems are made that confirm the theoretical estimates, and methods for dealing with boundary layer problems are illustrated.

Keywords: Elliptic second-order partial differential equations; Non-self-adjoint; Nonconforming finite element methods; Primal hybrid method; Cell discretization; Convection–diffusion equation

AMS classification: primary 65N30; 65M15 secondary 65N12

1. Introduction

A second-order linear partial differential operator is of the form

$$Au \equiv \sum_{i,j=1}^{K} b_{ij} D_i D_j u + \sum_{i=1}^{K} b_i D_i u + a_0 u,$$

where D_i denotes partial differentiation with respect to x_i. By absorbing certain derivatives of the functions $b_{ij}(x)$ into the first-order terms $\sum_{i=1}^{K} b_i D_i u$, we can convert this operator to the standard form

$$Au \equiv -\sum_{i,j=1}^{K} D_i(a_{ij} D_j u) + \sum_{i=1}^{K} a_i D_i u + a_0 u$$

* Corresponding author. E-mail: swann@sjsumcs.sjsu.edu.

PII S0377-0427(97)00051-4

by defining appropriate a_{ij} that will satisfy $a_{ij}(x) = a_{ji}(x)$. The operator is *elliptic* if there exists $c > 0$ such that

$$\sum_{i,j}^{K} a_{ij}(x) z_i z_j \geqslant c \sum_{i=1}^{K} z_i^2$$

for x in domain Ω for any $z_i \in \mathbb{R}$. The operator is said to be *self-adjoint* if the first-order terms $\sum_{i=1}^{K} b_i D_i u$ are absent. The standard example for a self-adjoint elliptic problem is the Poisson equation

$$-\Delta u = f.$$

Self-adjoint elliptic problems can be converted to a variational setting, where a certain functional is to be minimized. When applied to an approximation space consisting of finite linear combinations of basis functions, the functional induces a quadratic form and standard procedures for finding a minimum of a quadratic form produce a system of linear equations whose solutions provide the coefficients for the approximation. However, such variational methods do not readily apply to non-self-adjoint equations; thus some other method is more appropriate.

Our results hold for general non-self-adjoint second-order elliptic equations; the standard example is given by the steady-state convection–diffusion equations

$$\nabla \cdot (-\sigma \nabla u + u\boldsymbol{a}) = 0,$$

which describe the concentration $u(x)$ of a substance at point x, where the fluid that carries the substance is subject to a steady-state flow with velocity $\boldsymbol{a}(x)$ and σ is the diffusion coefficient.

We use a nonconforming extension of the finite element method called the *cell discretization algorithm* by Greenstadt [7–10]. It generalizes the *primal hybrid method* of Raviart and Thomas [14]. A domain is partitioned into *cells*; approximate solutions are given by linear combinations of functions that are part of any Schauder basis on each cell. We enforce a form of weak continuity on approximations over the entire domain by requiring that the difference of the traces of approximations on the common boundaries of adjacent cells be orthogonal to selected functions that are part of a basis defined on the interfaces between cells. These requirements, called *moment collocations*, give a set of linear constraints on the coefficients of the basis functions on each cell. For self-adjoint problems, the usual variational formulation allows us to use Lagrange multipliers to enforce these collocation constraints [4, 17]. This approach is not possible here, but we can use methods for solving parabolic equations [18] to construct a basis that carries with it the weak continuity constraints and then follows the standard Galerkin approximation procedure.

We describe the algorithm in Section 2 and, applying the results in [4, 17], we obtain error estimates that show convergence of approximations in H^1 on each cell to solutions of a homogeneous boundary value problems as the number of moment collocations enforced becomes large and the number of basis functions utilized becomes suitably larger.

In Section 3 we give methods for nonhomogeneous problems and convert our general error estimates to a polynomial implementation of the algorithm for domains partitioned into triangles and parallelograms. We express our error estimates in terms of an "h–p" finite element context [2, 3, 12]. We have created appropriate software that can generate the continuous approximations of a finite element method for comparison with our nonconforming approximations.

Our methods provide convergent approximations when applied to steady-state convection–diffusion equations when the underlying flow is constant or incompressible. Of particular interest are convection–dominated convection–diffusion equations. Two examples are given. A good approximation is obtained in the first example from [15]. The second example similar to those in [13] shows how a good choice of partition can assist in dealing with a boundary layer.

2. Convergence results

The setting for the method is described in full in [17, 18]. We summarize here.

We assume that bounded domain Ω in $\mathbb{R}^K$ (with boundary Γ) is partitioned into N subdomains Ω_i with Lipschitz continuous boundaries that are piecewise C^1; such subdomains are called *cells*. The exterior is $\Omega_0 \equiv \mathbb{R}^K \setminus \overline{\Omega}$.

Let $H^1(\Omega_i)$ denote the usual Hilbert space on each cell Ω_i with inner product $(\cdot,\cdot)_{1,i}$. The $H^1(\Omega_i)$ norm is denoted $\|\cdot\|_{1,i}$. $(\cdot,\cdot)_i$ represents the $L_2(\Omega_i)$ inner product, with the norm expressed as $\|\cdot\|_{0,i}$. $(\cdot,\cdot)$ denotes the $L_2(\Omega)$ inner product.

These spaces are assembled to form Hilbert space

$$H \equiv \{u \in L_2(\Omega)\colon u|_{\Omega_i} \in H^1(\Omega_i);\ i = 1,\ldots,N\}$$

with inner product

$$(u,v)_H \equiv \sum_{i=1}^{N} (u,v)_{1,i}$$

and norm represented by $\|\cdot\|_H$.

Let Γ_{ij} represent $\overline{\Omega}_i \cap \overline{\Omega}_j$. Γ_{i0} is a boundary segment between Ω_i and Ω_0. The inner product for $L_2(\Gamma_{ij})$ is denoted by $\langle\cdot,\cdot\rangle_{ij}$, with norm represented as $\|\cdot\|_{ij}$.

We denote by γ_{ij} the trace operator restricting $u|_{\Omega_i}$ to its values on Γ_{ij}. There are constants C_{ij} that depend on the geometry of Ω_i such that for any $u \in H$, $\|\gamma_{ij}(u)\|_{ij} \leqslant C_{ij}\|u\|_{1,i}$.

For each Γ_{ij}, choose $\{\omega_q^{ij}\}_{q=1}^{\infty}$ to be functions in $H^{1/2}(\Gamma_{ij})$ that are a Schauder basis for $L_2(\Gamma_{ij})$. Thus, for any $h \in L_2(\Gamma_{ij})$, there are coefficients d_k such that $h = \sum_{k=1}^{\infty} d_k \omega_k^{ij}$. Let $\mathscr{T}_n^{ij}(h) \equiv \sum_{k=n+1}^{\infty} d_k \omega_k^{ij}$. For any h and $\varepsilon > 0$, there is some $\mathrm{N}(h,\varepsilon)$ such that $n > \mathrm{N}(h,\varepsilon) \Rightarrow \|\mathscr{T}_n^{ij}(h)\|_{ij} < \varepsilon$.

For any $u \in H$, we define the kth *moment* of $u|_{\Omega_i}$ on Γ_{ij} to be

$$M_k^{ij}(u) \equiv \langle \gamma_{ij}(u), \omega_k^{ij} \rangle_{ij}.$$

Let N_I be the number of interfaces Γ_{ij}. Form multi-index $[n]$, an N_I-vector of nonnegative integers $(\ldots, n_{ij}, \ldots)$, with integer n_{ij} associated with interface Γ_{ij}.

Let $G[n] \equiv \{u \in H$: for any ij, $j \neq 0, ij = 1,\ldots,N_I$ and for any $k \leqslant n_{ij}$, we have $M_k^{ij}(u) = M_k^{ji}(u)\}$; this is the set of functions u in H such that the difference of the traces from either side of any internal interface Γ_{ij}, $\gamma_{ij}(u) - \gamma_{ji}(u)$, is $L_2(\Gamma_{ij})$-orthogonal to ω_k^{ij}, $k = 1,\ldots,n_{ij}$. We call such weak continuity across interfaces *moment collocation.*

Let $G_0[n] \equiv \{u \in G[n]$: for any i, for any $k \leqslant n_{i0}$, $M_k^{i0}(u) = 0\}$. $G_0[n]$ is the set of functions in $G[n]$ that are weakly zero on the external interfaces Γ_{i0} making up Γ.

We define a partial order for such multi-indices; we say $[n'] \geqslant [n] \Leftrightarrow$ for any ij, $n'_{ij} \geqslant n_{ij}$. If $[n^k]$ is a sequence of multi-indices, $k = 1, 2, \ldots$, we say that $[n^k] \to [\infty]$ if $[n^k] \leqslant [n^{k+1}]$ and $\inf_{ij}\{n^k_{ij}\} \to \infty$ as $k \to \infty$.

For the ith cell Ω_i, choose any Schauder basis $\{B^i_k\}$ for $H^1(\Omega_i)$. Thus, for any v in $H^1(\Omega_i)$, there are b^i_k such that $\sum_{k=1}^{\infty} b^i_k B^i_k = v$; let $v_{\cdot,m} = \sum_{k=1}^{m} b^i_k B^i_k$. Let $\mathscr{Q}^i_m(v)$ denote the orthogonal projection (in the $H^1(\Omega_i)$ inner product) of v onto the $H^1(\Omega_i)$-orthogonal *complement* of the span of $\{B^i_1, B^i_2, \ldots, B^i_m\}$. Thus

$$\mathscr{Q}^i_m(v_{\cdot,m}) = 0; \qquad \mathscr{Q}^i_m(v) = \mathscr{Q}^i_m(v - v_{\cdot,m});$$

$$\|\mathscr{Q}^i_m(v)\|_{1,i} \leqslant \|v - v_{\cdot,m}\|_{1,i} = \left\| \sum_{k=m+1}^{\infty} b^i_k B^i_k \right\|_{1,i} \quad \text{and} \quad \lim_{m\to\infty} \|\mathscr{Q}^i_m(v)\|_{1,i} = 0.$$

Let $[m]$ be an N-dimensional multi-index indicating the number of basis functions used in the approximation on each of the N cells; we adopt the same notational conventions as those used for multi-index $[n]$.

Define $H[m]$ to be the subspace of H such that for any $v \in H[m]$, $v|_{\Omega_i}$ is in the span of $\{B^i_1, B^i_2, \ldots, B^i_{m_i}\}$.

Given $[m]$, and any function v in H, $\mathscr{Q}_{[m]}(v)$ is defined to be the function in H such that $\mathscr{Q}_{[m]}(v)|_{\Omega_i} = \mathscr{Q}^i_{m_i}(v|_{\Omega_i})$. Thus, $\mathscr{Q}_{[m]}(\cdot)$ is the projection of H onto $H[m]^{\perp}$; $\lim_{[m]\to[\infty]} \|\mathscr{Q}_{[m]}(v)\|_H = 0$.

Let $G_0[n][m] \equiv G_0[n] \cap H[m]$. The moment collocation requirements are met by requiring that certain linear equations hold among the b^i_k, e.g., for $u \in G_0[n][m]$, we require that, on internal interfaces Γ_{ij},

$$\langle \gamma_{ij}(u), \omega^{ij}_p \rangle_{ij} - \langle \gamma_{ji}(u), \omega^{ij}_p \rangle_{ij} = 0, \quad p = 1, \ldots, n_{ij},$$

which gives the requirement

$$\sum_{k=1}^{m_i} b^i_k \langle \gamma_{ij}(B^i_k), \omega^{ij}_p \rangle_{ij} - \sum_{k=1}^{m_j} b^j_k \langle \gamma_{ji}(B^j_k), \omega^{ij}_p \rangle_{ij} = 0 \tag{1}$$

and, for the external boundary segments Γ_{i0},

$$\langle \gamma_{i0}(u), \omega^{i0}_p \rangle_{i0} = 0, \quad p = 1, \ldots, n_{i0},$$

which becomes

$$\sum_{k=1}^{m_i} b^i_k \langle \gamma_{ij}(B^i_k), \omega^{i0}_p \rangle_{i0} = 0. \tag{2}$$

The following lemma connects estimates in terms of these spaces.

Lemma 1.1. *If $\mathscr{P}^n_m$ is the orthogonal projection operator of $G_0[n]$ onto $G_0[n][m]$, then there is a constant $K = K([n])$ depending on the number '$[n]$' of moment collocations, the cell decomposition of domain Ω and the choice of basis functions and collocation functions such that*

$$\|u - \mathscr{P}^n_m u\|_H \leqslant K([n]) \|\mathscr{Q}_{[m]} u\|_H,$$

where $\mathscr{Q}_{[m]}$ is the projection of H onto the H-orthogonal complement of $H[m]$.

The proof of this lemma is found in [17]; see [4] for a detailed discussion of – and estimates for – $K([n])$. We return to the consideration of $K([n])$ in the discussion of a polynomial implementation of the algorithm given in Section 3.

We construct a new basis for $G_0[n][m]$ that carries with it the moment collocation constraints.

The coefficients $\{b_k^i\}$ for the representation on each of the N cells can be concatenated to form vector $\boldsymbol{b} \equiv (b_1^1, b_2^1, \ldots, b_1^2, b_2^2, \ldots, b_1^k, b_2^k, \ldots)$. The linear moment collocation requirements (1) and (2) are expressed as $\boldsymbol{M}\boldsymbol{b}^{\mathrm{T}} = \boldsymbol{0}$, for a suitable rectangular matrix $\boldsymbol{M}$; it is an $n' \times m$ matrix, where $n' = \sum n_{ij}$ and $m = \sum_{k=1}^{N} m_k$; $m > n'$. It is shown in [17] that, for any $[n]$, the rows of $\boldsymbol{M}$ are independent if $[m]$ is sufficiently large. Dorr [6] discusses this phenomena using polynomial approximations in an *hp* setting; we encounter row dependency in the experiments described in Section 3 and describe appropriate increases in $[m]$.

The set of acceptable arrays of coefficients $\boldsymbol{b}$ that can be used to define functions in $G_0[n][m]$ is the null space of $\boldsymbol{M}$, which we obtain as follows:

Compute the "QR" factorization of $\boldsymbol{M}^{\mathrm{T}}$, so $\boldsymbol{M}^{\mathrm{T}} = (\boldsymbol{Q}'' \mid \boldsymbol{Q}') \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix}$, where $\boldsymbol{R}$ is square upper-triangular and invertible and $Q \equiv (\boldsymbol{Q}'' \mid \boldsymbol{Q}')$ is orthogonal. Then $\boldsymbol{M} = R^{\mathrm{T}} Q^{\mathrm{T}}$, where $R \equiv \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix}$. Since we are looking for $\boldsymbol{b}$ such that $\boldsymbol{M}\boldsymbol{b}^{\mathrm{T}} = \boldsymbol{0}$, a straightforward argument shows that the columns of $\boldsymbol{Q}'$, the last $m - n'$ columns of Q, are an orthonormal basis for the null space of $\boldsymbol{M}$.

Let $p \equiv m - n'$ and suppose that the p columns of $\boldsymbol{Q}'$ are $(q_{11}, \ldots, q_{m1})^{\mathrm{T}}, (q_{12}, \ldots, q_{m2})^{\mathrm{T}}, \ldots, (q_{1p}, \ldots, q_{mp})^{\mathrm{T}}$.

We enumerate the $\{B_k^i\}$ as

$$\{B_1^1, B_2^1, \ldots, B_{m_1}^1, B_1^2, B_2^2, \ldots, B_{m_2}^2, \ldots\};$$

there are m such B_k^i. Denote the B_k^i with this enumeration as $\{\Phi^1, \Phi^2, \ldots, \Phi^m\}$ and form $\mathscr{B}_i \equiv \sum_{j=1}^{m} q_{ji} \Phi^j$ defined on all of Ω by assuming each B_k^j is zero outside Ω_j. Then $\{\mathscr{B}_i\}$ is a basis for $G_0[n][m]$. Any approximation of form $u_{n,m} = \sum_{i=1}^{p} \beta_i \mathscr{B}_i$ can be expressed in terms of the original basis represented by $\{\Phi^1, \Phi^2, \ldots, \Phi^m\}$ using coefficients ϕ_j, the components of vector $\phi = \boldsymbol{Q}'\boldsymbol{y}^{\mathrm{T}}$, where $y = (\beta_1, \beta_2, \ldots, \beta_p)$.

We consider the following non-self-adjoint problem: Let A be the linear operator

$$Au \equiv -\sum_{i,j=1}^{K} D_i(a_{ij} D_j u) + \sum_{i=1}^{K} a_i D_i u + a_0 u.$$

We consider the homogeneous Dirichlet problem $Au = f$ with $\gamma(u) = 0$.

We treat elliptic problems, so we assume that the operator

$$Eu \equiv -\sum_{i,j=1}^{K} D_i(a_{ij}(x) D_j u) + a_0 u$$

is elliptic, self-adjoint and coercive. It suffices that the following hold:

(*E*1) All a_{ij} and a_0 are in $C^1(\overline{\Omega})$,

(*E*2) There exists $c > 0$ such that

$$\sum_{i,j}^{K} a_{ij}(x) z_i z_j \geqslant c \sum_{i=1}^{K} z_i^2$$

for x in Ω and for any $z_i \in \mathbb{R}$ and $a_{ij}(x) = a_{ji}(x)$ and
(E3) $a_0(x) > 0$ for $x \in \Omega$.

We assume that the functions a_i forming the first-order term $\sum_{i=1}^{K} a_i D_i u$, denoted by $\boldsymbol{a} \cdot \nabla u$, are measurable and bounded.

Operator E induces a coercive bilinear form

$$a(u, v) \equiv \int_{\Omega} \sum_{i,j}^{K} a_{ij}(x) D_i u D_j v + a_0(x) uv \, \mathrm{d}x.$$

There is a constant depending on the a_{ij} and a_0 such that for any $v \in H, a(v, v) \leqslant M \|v\|_H^2$.

Green's formula relates E and $a(\cdot, \cdot)$.

Let $D_{n_{ij}} u$ be the "co-normal derivative with respect to E of u on Γ_{ij}" defined as follows: If $\boldsymbol{n} = (n_1, n_2, \ldots, n_K)$ is the unit normal to Γ_{ij} (pointing outward relative to the the interior of Ω_i), then

$$D_{n_{ij}} u \equiv \sum_{p,q}^{K} \gamma_{ij} (a_{pq} D_q u) n_p.$$

Let $D_n u$ denote the general co-normal derivative. Green's formula

$$(\boldsymbol{E}u, v) = a(u, v) - \langle D_n u, \gamma(v) \rangle_{\Gamma'} \tag{3}$$

holds for suitable domains Ω' with boundary Γ' for u in $H^2(\Omega')$ and v in $H^1(\Omega')$ if the a_{ij} are sufficiently smooth [19]. In particular, Green's formula is valid with our assumptions concerning the a_{ij} and $\Omega' = \Omega$ or $\Omega' =$ any Ω_j.

The trace of the solution u is to be zero on $\partial\Omega$. The weak form of the problem is the following: For solution $u \in H_0^1(\Omega)$, we require that

$$a(u, v) + (\boldsymbol{a} \cdot \nabla u, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega).$$

To obtain an approximation, we use the Galerkin method with the basis $\{\mathscr{B}_i\}$ for the space $G_0[n][m]$. An approximation is of form $u_p \equiv \sum_{i=1}^{p} \beta_i \mathscr{B}_i(x)$, and we require that

$$a(u_p, \mathscr{B}_j) + (\boldsymbol{a} \cdot \nabla u_p, \mathscr{B}_j) = (f, \mathscr{B}_j) \quad \text{for } j = 1, \ldots, p.$$

We express this in terms of the original basis $\{B_i^k\}$ as follows: Let $\boldsymbol{C}$ denote the matrix of positive definite diagonal blocks $(a(B_i^k, B_j^k))$, so $a(\mathscr{B}_i, \mathscr{B}_j) = \boldsymbol{Q}'^{\mathrm{T}} \boldsymbol{C} \boldsymbol{Q}'$. If we denote the (nonsymmetric) matrix of diagonal blocks $((\boldsymbol{a} \cdot \nabla B_r^k, B_s^k))$ by $\boldsymbol{H}$, $((\boldsymbol{a} \cdot \nabla \mathscr{B}_i, \mathscr{B}_j)) = \boldsymbol{Q}'^{\mathrm{T}} \boldsymbol{H} \boldsymbol{Q}'$. Finally, $(f, \mathscr{B}_j) = ((f, B_i^k)) \boldsymbol{Q}'$.

We have represented the row of undetermined coefficients by β; we obtain our approximation by solving

$$\beta \boldsymbol{Q}'^{\mathrm{T}} (\boldsymbol{C} + \boldsymbol{H}) \boldsymbol{Q}' = ((f, B_i^k)) \boldsymbol{Q}'. \tag{4}$$

The size of the matrix $\boldsymbol{Q}'^{\mathrm{T}} (\boldsymbol{C} + \boldsymbol{H}) \boldsymbol{Q}'$ is $p \times p$, where
$p =$ (size of matrix $\boldsymbol{C}$) − (total number of collocations enforced).

For a problem with 16 rectangular cells, using the full basis for the set of polynomials of degree 10 or less and 8 collocations on each Γ_{ij}, p is about 736. The p degrees of freedom in the system pertain to the approximation of the solution of the equation; we have eliminated any concern with weak continuity between interfaces of cells.

We give two convergence results.

If we assume that the problem is *elliptic* and *coercive* in the sense of Wloka [19], then there exist positive constants c_1 and c_2, such that for any u and v in $G_0[n]$,
(a) $|a(u,v)+(\boldsymbol{a}\cdot\nabla u,v)|\leqslant c_1\|u\|_H\|v\|_H$ and
(b) $a(u,u)+(\boldsymbol{a}\cdot\nabla u,u)\geqslant c_2\|u\|_H^2$.

The coercivity in (b) implies that the linear system has a unique solution β, which we use to define an approximate solution $u_{n,m}\equiv\sum_{i=1}^{p}\beta_i\mathscr{B}_i(x)$. Our first theorem makes the assumption that (a) and (b) hold. The second theorem assumes only that the system of linear equations (4) has a unique solution.

Theorem 1.2. *Suppose that operator E satisfies (E1)–(E3), functions a_i are bounded and measurable, $f\in L_2(\Omega)$, and that u is a solution in $H_0^2(\Omega)$ to $Eu+\boldsymbol{a}\cdot\nabla u=f$. Suppose that there is some $c_2>0$, such that for all $v\in G_0[n]$, $a(v,v)+(\boldsymbol{a}\cdot\nabla v,v)\geqslant c_2\|v\|_H^2$. Let $M_1\equiv\sqrt{K}\sup\{|a_i(x)|\}$, where the supremum is taken over all $x\in\Omega$ and $1\leqslant i\leqslant K$. If $u_{n,m}$ is the approximation obtained by solving the system of linear equations, then*

$$c_2\|u-u_{n,m}\|_H\leqslant(M+M_1)K([n])\|\mathscr{Q}_{[m]}u\|_H+n_f\sqrt{N}\sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\boldsymbol{n}_{ij}}u)\|_{ij}\},$$

where $K([n])$ is the parameter of Lemma 1.1 and n_f is the maximum number of faces of any of the N cells.

Proof. The Galerkin approximation satisfies the relation

$$a(u_{n,m},v)+(\boldsymbol{a}\cdot\nabla u_{n,m},v)=(f,v)\text{ for all }v\in G_0[n][m].$$

Let $\mathscr{P}_m^n v$ denote the projection of $v\in G_0[n]$ onto $G_0[n][m]$.

$$\begin{aligned}
&a(u-u_{n,m},u-u_{n,m})+(\boldsymbol{a}\cdot\nabla(u-u_{n,m}),u-u_{n,m})\\
&\quad=a(u-u_{n,m},u-\mathscr{P}_m^n u)+(\boldsymbol{a}\cdot\nabla(u-u_{n,m}),u-\mathscr{P}_m^n u)\\
&\qquad+a(u-u_{n,m},\mathscr{P}_m^n u-u_{n,m})+(\boldsymbol{a}\cdot\nabla(u-u_{n,m}),\mathscr{P}_m^n u-u_{n,m})\\
&\quad=a(u-u_{n,m},u-\mathscr{P}_m^n u)+(\boldsymbol{a}\cdot\nabla(u-u_{n,m}),u-\mathscr{P}_m^n u)\\
&\qquad+a(u,\mathscr{P}_m^n u-u_{n,m})+(\boldsymbol{a}\cdot\nabla u,\mathscr{P}_m^n u-u_{n,m})\\
&\qquad-a(u_{n,m},\mathscr{P}_m^n u-u_{n,m})-(\boldsymbol{a}\cdot\nabla u_{n,m},\mathscr{P}_m^n u-u_{n,m}).
\end{aligned}$$

Let $\delta=\mathscr{P}_m^n u-u_{n,m}=\mathscr{P}_m^n(u-u_{n,m})\in G_0[n][m]$. Then

$$\begin{aligned}
&a(u_{n,m},\mathscr{P}_m^n u-u_{n,m})+(\boldsymbol{a}\cdot\nabla u_{n,m},\mathscr{P}_m^n u-u_{n,m})\\
&\quad=a(u_{n,m},\delta)+(\boldsymbol{a}\cdot\nabla u_{n,m},\delta)=(f,\delta).
\end{aligned}$$

As in [17], we use Green's formula to get

$$\begin{aligned}
&a(u,\mathscr{P}_m^n u-u_{n,m})+(\boldsymbol{a}\cdot\nabla u,\mathscr{P}_m^n u-u_{n,m})=a(u,\delta)+(\boldsymbol{a}\cdot\nabla u,\delta)\\
&\quad=(Eu,\delta)+(\boldsymbol{a}\cdot\nabla u,\delta)+\sum_{\Gamma_{ij}}\langle D_{\boldsymbol{n}_{ij}}u,\gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij}+\sum_{\Gamma_{i0}}\langle D_{\boldsymbol{n}_{i0}}u,\gamma_{i0}(\delta)\rangle_{i0}\\
&\quad=(f,\delta)+\sum_{\Gamma_{ij}}\langle D_{\boldsymbol{n}_{ij}}u,\gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij}+\sum_{\Gamma_{i0}}\langle D_{\boldsymbol{n}_{i0}}u,\gamma_{i0}(\delta)\rangle_{i0}.
\end{aligned}$$

Using the argument in [17], since δ is in $G_0[n][m]$, the differences of the traces of δ on any Γ_{ij} are orthogonal to the $[n]$ collocation functions and we can use Schwarz' inequality to obtain estimate

$$\left|\sum_{\Gamma_{ij}}\langle D_{\boldsymbol{n}_{ij}}u,\gamma_{ij}(\delta)-\gamma_{ji}(\delta)\rangle_{ij}+\sum_{\Gamma_{i0}}\langle D_{\boldsymbol{n}_{i0}}u,\gamma_{i0}(\delta)\rangle_{i0}\right|$$

$$\leqslant \sup\{C_{ij}\}\sup\{\|\mathscr{T}^{ij}_{n_{ij}}(D_{\boldsymbol{n}_{ij}}u)\|_{ij}\}(n_f)\sqrt{N}\|\delta\|_H,$$

where C_{ij} are the trace constants, $\|\mathscr{T}^{ij}_{n_{ij}}(D_{\boldsymbol{n}_{ij}}u)\|_{ij}$ is the norm of that part of the trace of the normal derivative of u on Γ_{ij} that is *not* in the span of the moment collocation functions, n_f is the maximum number of faces of any cell, and N is the number of cells. Recalling the definition of δ we have

$$\|\delta\|_H = \|\mathscr{P}^n_m(u-u_{n,m})\|_H \leqslant \|u-u_{n,m}\|_H.$$

For any $u, v \in H$, the Schwarz inequality gives the estimate

$$|(\boldsymbol{a}\cdot\nabla u, v)| \leqslant \sup\{|a_i(x)|\}\sum_{i=1}^{K}\|D_i u\|_0\|v\|_0 \leqslant \sup\{|a_i(x)|\}k^{1/2}\|u\|_H\|v\|_0.$$

We have defined M_1 to be $\sup\{|a_i(x)|\}k^{1/2}$. Recall that $|a(u,v)| \leqslant M\|u\|_H\|v\|_H$. These estimates give

$$\begin{aligned}
&a(u-u_{n,m},u-u_{n,m})+(\boldsymbol{a}\cdot\nabla(u-u_{n,m}),u-u_{n,m})\\
&\quad\leqslant M\|u-u_{n,m}\|_H\|u-\mathscr{P}^n_m u\|_H + M_1\|u-u_{n,m}\|_H\|u-\mathscr{P}^n_m u\|_0\\
&\qquad+(f,\delta)+\sup\{C_{ij}\}\sup\{\|\mathscr{T}^{ij}_{n_{ij}}(D_{\boldsymbol{n}_{ij}}u)\|_{ij}\}(n_f)\sqrt{N}\|u-u_{n,m}\|_H-(f,\delta)\\
&\quad\leqslant \|u-u_{n,m}\|_H\{(M+M_1)\|u-\mathscr{P}^n_m u\|_H+\sup\{C_{ij}\}\sup\{\|\mathscr{T}^{ij}_{n_{ij}}(D_{\boldsymbol{n}_{ij}}u)\|_{ij}\}n_f\sqrt{N}\}. \qquad (5)
\end{aligned}$$

By assumption

$$c_2\|u-u_{n,m}\|^2_H \leqslant a(u-u_{n,m},u-u_{n,m})+(\boldsymbol{a}\cdot\nabla(u-u_{n,m}),u-u_{n,m}),$$

so term $\|u-u_{n,m}\|_H$ can be canceled from both sides of the inequality above.

Since $\|u-\mathscr{P}^n_m u\|_H \leqslant K([n])\|\mathscr{Q}_{[m]}u\|_H$, we obtain the estimate of the theorem. □

The coercivity condition $a(u,u)+(\boldsymbol{a}\cdot\nabla u,u)\geqslant c_2\|u\|^2_H$ is stronger than necessary to insure existence of a unique solution [11, 19]. Grisvard [11], for example, shows that it suffices that $a_0(x)\geqslant c_5>0$ and that the a_i be bounded and measurable. Our next theorem gives a result for approximations obtained under the assumption that there is a unique solution $u_{n,m}$ to (4) and only form $a(u,v)$ is coercive.

Theorem 1.3. *Suppose that form $a(u,v)$ satisfies the requirements of Theorem* 1.2 *and functions a_i are in $C^1(\bar{\Omega})$ and there is some $c_3>0$ such that*

$$a_0(x)-(\tfrac{1}{2})\nabla\cdot\boldsymbol{a}(x)\geqslant c_3$$

for all x in Ω. Then there is a constant $c_4>0$ such that

$$c_4\|u-u_{n,m}\|_H$$
$$\leqslant (M+M_1)K([n])\|\mathcal{Q}_{[m]}u\|_H+\sqrt{N}n_f\sup\{C_{ij}\}\sup\{\|\mathcal{T}_{n_{ij}}^{ij}(D_{\boldsymbol{n}_{ij}}u)\|_{ij}\}$$
$$+(\tfrac{1}{2})\sqrt{n_f}\sup\{C_{ij}\}M_1\left\{\sum_{\Gamma_{ij}}\|\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})\|_{ij}^2+\sum_{\Gamma_{i0}}\|\gamma_{i0}(u_{n,m})\|_{i0}^2\right\}^{1/2}.$$

Proof. For any $v\in H$,

$$(\boldsymbol{a}\cdot\nabla v,v)=\int_\Omega\left(\sum_{i=1}^K a_iD_iv\right)v\,\mathrm{d}x=\frac{1}{2}\int_\Omega\sum_{i=1}^K a_iD_iv^2\,\mathrm{d}x$$
$$=\frac{1}{2}\int_\Omega\nabla\cdot(v^2\boldsymbol{a})\,\mathrm{d}x-\frac{1}{2}\int_\Omega v^2\nabla\cdot\boldsymbol{a}\,\mathrm{d}x.$$

By the divergence theorem

$$\int_\Omega\nabla\cdot(v^2\boldsymbol{a})\,\mathrm{d}x=\sum_{i=1}^N\int_{\Omega_i}\nabla\cdot(v^2\boldsymbol{a})\,\mathrm{d}x=\sum_{i=1}^N\left(\sum_j\int_{\Gamma_{ij}}\gamma_{ij}(v^2\boldsymbol{a})\cdot\boldsymbol{n}_{ij}\,\mathrm{d}s\right),$$

where $\boldsymbol{n}_{ij}$ is the unit outward normal to Γ_{ij}.

Using the fact that if the boundary segment is an internal interface, $\gamma_{ij}(\boldsymbol{a})\cdot\boldsymbol{n}_{ij}=-\gamma_{ji}(\boldsymbol{a})\cdot\boldsymbol{n}_{ji}$, and grouping the sum of the boundary integrals above in pairs (if the boundary segment is an internal interface), we obtain

$$\sum_{i=1}^N\left(\sum_j\int_{\Gamma_{ij}}\gamma_{ij}(v^2\boldsymbol{a})\cdot\boldsymbol{n}_{ij}\,\mathrm{d}s\right)=\sum_{\Gamma_{ij}}\int_{\Gamma_{ij}}[\gamma_{ij}(v^2)-\gamma_{ji}(v^2)]\gamma_{ij}(\boldsymbol{a})\cdot\boldsymbol{n}_{ij}\,\mathrm{d}s$$
$$+\sum_{\Gamma_{i0}}\int_{\Gamma_{i0}}\gamma_{i0}(v^2)\gamma_{i0}(\boldsymbol{a})\cdot\boldsymbol{n}_{i0}\,\mathrm{d}s.$$

The first sum in the expression above is taken over $j\neq0$ and we assume that $i<j$.

We let $v=u-u_{n,m}$. For internal interfaces Γ_{ij},

$$\gamma_{ij}((u-u_{n,m})^2)-\gamma_{ji}((u-u_{n,m})^2)$$
$$=\gamma_{ij}(u^2-2uu_{n,m}+u_{n,m}^2)-\gamma_{ji}(u^2-2uu_{n,m}+u_{n,m}^2)$$
$$=\gamma_{ij}(u^2)-\gamma_{ji}(u^2)-2\gamma_{ij}(u)[\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})]+[\gamma_{ij}(u_{n,m})]^2-[\gamma_{ji}(u_{n,m})]^2.$$

Since

$$\gamma_{ij}(u)=\gamma_{ji}(u),\gamma_{ij}((u-u_{n,m})^2)-\gamma_{ji}((u-u_{n,m})^2)$$
$$=[\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})][-2\gamma_{ij}(u)+\gamma_{ij}(u_{n,m})+\gamma_{ji}(u_{n,m})]$$
$$=[\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})][\gamma_{ij}(u_{n,m}-u)+\gamma_{ji}(u_{n,m}-u)].$$

The homogeneous boundary condition gives $\gamma_{i0}(u)=0$, so

$$\gamma_{i0}((u-u_{n,m})^2)=\gamma_{i0}((u_{n,m})^2)=\gamma_{i0}(u_{n,m})\gamma_{i0}(u_{n,m}-u).$$

Assembling these results, we get

$$\begin{aligned}
&2(\boldsymbol{a}\cdot\nabla(u-u_{n,m}),u-u_{n,m})\\
&\quad=-\int_\Omega(u-u_{n,m})^2\nabla\cdot\boldsymbol{a}\,\mathrm{d}x+\sum_{\Gamma_{ij}}\int_{\Gamma_{ij}}[\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})][\gamma_{ij}(u_{n,m}-u)\\
&\qquad+\gamma_{ji}(u_{n,m}-u)]\gamma_{ij}(\boldsymbol{a})\cdot\boldsymbol{n}_{ij}\,\mathrm{d}s+\sum_{\Gamma_{i0}}\int_{\Gamma_{i0}}\gamma_{i0}(u_{n,m})\gamma_{i0}(u_{n,m}-u)\gamma_{i0}(\boldsymbol{a})\cdot\boldsymbol{n}_{i0}\,\mathrm{d}s.
\end{aligned}$$

The use of Schwarz' inequality readily shows that constant M_1 of Theorem 1.2 majorizes the supremum of $|\gamma_{ij}(\boldsymbol{a})\cdot\boldsymbol{n}_{ij}|$ over all Γ_{ij} and Γ_{i0}. Then

$$\begin{aligned}
&\left|\int_{\Gamma_{ij}}[\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})][\gamma_{ij}(u_{n,m}-u)+\gamma_{ji}(u_{n,m}-u)]\gamma_{ij}(\boldsymbol{a})\cdot\boldsymbol{n}_{ij}\,\mathrm{d}s\right|\\
&\quad\leqslant M_1\|\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})\|_{ij}[\|\gamma_{ij}(u_{n,m}-u)\|_{ij}+\|\gamma_{ji}(u_{n,m}-u)\|_{ji}]\\
&\quad\leqslant M_1\|\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})\|_{ij}[C_{ij}\|u_{n,m}-u\|_{1,i}+C_{ji}\|u_{n,m}-u\|_{1,j}].
\end{aligned}$$

Likewise

$$\left|\int_{\Gamma_{i0}}\gamma_{i0}(u_{n,m})\gamma_{i0}(u_{n,m}-u)\gamma_{i0}(\boldsymbol{a})\cdot\boldsymbol{n}_{i0}\,\mathrm{d}s\right|\leqslant M_1\|\gamma_{i0}(u_{n,m})\|_{i0}C_{i0}\|u_{n,m}-u\|_{1,i}.$$

Using the Schwarz inequality, and the fact that any $\|u_{n,m}-u\|_{1,i}$ occurs at most n_f times in the sums, the sums over the Γ_{ij} and Γ_{i0} are majorized by

$$M_1\left\{\sum_{\Gamma_{ij}}\|\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})\|_{ij}^2+\sum_{\Gamma_{i0}}\|\gamma_{i0}(u_{n,m})\|_{i0}^2\right\}^{1/2}\sup\{C_{ij}\}(n_f)^{1/2}\left\{\sum_{i=1}^{N}\|u_{n,m}-u\|_{1,i}^2\right\}^{1/2}.$$

This last sum is $\|u_{n,m}-u\|_H$. We combine these estimates with inequality (5) of Theorem 1.2:

$$\begin{aligned}
&a(u-u_{n,m},u-u_{n,m})-\frac{1}{2}\int_\Omega(u-u_{n,m})^2\nabla\cdot\boldsymbol{a}\,\mathrm{d}x\\
&\quad\leqslant\|u-u_{n,m}\|_H\left[(M+M_1)\|u-\mathscr{P}_m^n u\|_H+\sup\{C_{ij}\}\sup\{\|\mathscr{T}_{n_{ij}}^{ij}(D_{\boldsymbol{n}_{ij}}u)\|_{ij}\}n_f\sqrt{N}\right.\\
&\qquad\left.+\left(\frac{1}{2}\right)\sup\{C_{ij}\}(n_f)^{1/2}M_1\left\{\sum_{\Gamma_{ij}}\|\gamma_{ij}(u_{n,m})-\gamma_{ji}(u_{n,m})\|_{ij}^2+\sum_{\Gamma_{i0}}|\gamma_{i0}(u_{n,m})\|_{i0}^2\right\}^{1/2}\right]
\end{aligned}\tag{6}$$

The supposition concerning c_3 means that

$$a(u-u_{n,m},u-u_{n,m})-\frac{1}{2}\int_\Omega (u-u_{n,m})^2\nabla\cdot\boldsymbol{a}\,\mathrm{d}x$$
$$=\int_\Omega \sum_{i,j}^{K} a_{ij}(x)D_i(u-u_{n,m})D_j(u-u_{n,m})+\left[a_0(x)-\left(\frac{1}{2}\right)\nabla\cdot\boldsymbol{a}\right](u-u_{n,m})^2\,\mathrm{d}x$$
$$\geqslant c_4\|u-u_{n,m}\|_H^2,$$

so the term $\|u-u_{n,m}\|_H$ can be canceled from each side of inequality (6) above. □

Due to the moment collocation constraints, the sums in the estimate over the Γ_{ij} and Γ_{i0} are small; more so as $[n]$ is large. Such interface error terms are independent of the solution u, and are readily computed in the solution process. In the polynomial implementation in Section 3 approximations can be made continuous, thus eliminating these error terms.

If the vector field $\boldsymbol{a}$ is constant or $\nabla\cdot\boldsymbol{a}\leqslant 0$, the assumptions concerning the existence of an appropriate constant c_3 are unnecessary. In fact, a stronger result holds in this case; the following corollary also implies that the estimates of Theorem 1.3 apply to the steady-state convection–diffusion equations

$$\nabla\cdot(-\sigma\nabla u+u\boldsymbol{a})=0$$

for concentration $u(x)$, when the underlying flow with velocity $\boldsymbol{a}$ is incompressible.

Corollary 1.4. *Suppose that form $a(u,v)$ satisfies the requirements of Theorem* 1.2 *and functions a_i are in $C^1(\bar{\Omega})$. If the vector field $\boldsymbol{a}$ satisfies $\nabla\cdot\boldsymbol{a}\leqslant 0$, and $[n]$ is sufficiently large so that for each Γ_{ij} there is some $k\leqslant n_{ij}$ such that $\langle\omega_k^{ij},1\rangle_{ij}\neq 0$, then coefficient $a_0(x)$ can be zero and the estimate of Theorem* 1.3 *is valid.*

Proof. From [4], when the additional assumption is made concerning the size of $[n]$, functional $a(u,v)$ is coercive over $G_0[n]$, and the term containing nonpositive $\nabla\cdot\boldsymbol{a}$ in Eq. (6) can be discarded. □

3. Nonhomogeneous boundary value problems

We adapt the algorithm to nonhomogeneous boundary value problems, where, given some $g(\cdot)$ defined on Γ, we now require that $u(x)=g(x)$. The classical method is to first find some $u_2(x)$ such that $u_2=g$ on Γ, next express u as u_1+u_2, with $u_1=0$ on Γ, and then solve $Au_1=f-Au_2$.

The equivalent transformed problem requires that we find some vector η such that $\boldsymbol{M}\eta=\boldsymbol{g}$, where the components of $\boldsymbol{g}$ corresponding to internal collocation on Γ_{ij} are zero; collocation rows corresponding to Γ_{i0} produce an entry of form $\langle g(\cdot),\omega_q^{i0}\rangle_{i0}$. An η of minimal norm satisfying these requirements can be obtained as part of the $\boldsymbol{QR}$ factorization of $\boldsymbol{M}^{\mathrm{T}}$: if $\boldsymbol{M}^{\mathrm{T}}=(\boldsymbol{Q}''|\boldsymbol{Q}')\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}$, where

$\boldsymbol{R}$ is square upper-triangular and invertible, then $\eta = \boldsymbol{Q}''(\boldsymbol{R}^{\mathrm{T}})^{-1}\boldsymbol{g}$. We approximate u_2 by

$$u_2|_{\Omega_k} = \sum_{i=1}^{m_k} \eta_i^k B_i^k,$$

where η_i^k is the component of η associated with B_i^k. Suppose that our approximation to u_1 expressed in terms of the original basis is

$$u_1|_{\Omega_k} = \sum_{i=1}^{m_k} b_i^k B_i^k.$$

If we take the $L_2(\Omega_k)$ inner product of $Au_1 = f - Au_2$ with B_i^k and follow the argument in Section 2, the resulting vector equation is

$$\boldsymbol{b}(\boldsymbol{C} + \boldsymbol{H}) = \boldsymbol{f} - \eta^{\mathrm{T}}(\boldsymbol{C} + \boldsymbol{H}),$$

where $\boldsymbol{f} \equiv ((f, B_i^k)_0)$ and $\boldsymbol{b}$ is to satisfy $\boldsymbol{M}\boldsymbol{b}^{\mathrm{T}} = \boldsymbol{0}$.

In terms of the basis $\{\mathscr{B}_j\}$ satisfying the collocation constraints, u_1 is represented as $\sum_i^p \beta_i \mathscr{B}_i(x)$ and vector β is to satisfy

$$\beta \boldsymbol{Q}'^{\mathrm{T}}(\boldsymbol{C} + \boldsymbol{H})\boldsymbol{Q}' = \boldsymbol{f}\boldsymbol{Q}' - \eta^{\mathrm{T}}(\boldsymbol{C} + \boldsymbol{H})\boldsymbol{Q}'.$$

We have implemented this scheme for arbitrary problems with domains in $\mathbb{R}^2$ that can be partitioned into triangles or parallelograms (or both). We use L_2-orthonormal bases of polynomials of degree 10 or less (up to 66 functions on each cell) to provide approximations. Gauss–Legendre quadrature is used to effect the integrations over the cells and interfaces, and subroutines from LINPACK [5] and LAPACK [1] provide the QR decomposition and the solution of the final system.

We test the error estimates in terms of p, the degree of the polynomial approximation on each cell, q, the maximum degree of the Legendre polynomials providing the weight functions on the interfaces in any trial, and h, the maximum diameter of the cells in the cell decomposition of Ω. We use the same number $q+1$ of collocations on each boundary segment Γ_{ij}, so we replace all n_{ij} with $q+1$; we revise the notation containing collocation index $[n]$ by replacing $[n]$ with q. We also use the same number of basis functions for our pth order basis on each cell; the notation containing basis multi-index $[m]$ replaces $[m]$ by p. Thus, we now denote approximation $u_{n,m}$ with symbol $u_{q,p}$. The relevant error estimates for a polynomial implementation of these methods are given in [4]. They are expressed in terms of $H^k(\Omega)$ norm of the solution ($k > 2$) or, for analytic solutions, in terms of the semi-norm defined by the L_2-norm of the $p+1$ and $p+2$ derivatives of the solution. Since our test problem is analytical we use these second estimates.

The following estimates hold for domains in $\mathbb{R}^2$:

The trace constants C_{ij} for boundaries of polygonal cells are bounded by $c_1/(h)^{1/2}$, where c_1 is independent of h and depends only on the smallest angle in any cell.

$K[n]$ is bounded by $c_2(q)/h$. We can get some sense of the size of $c_2(q)$ from the experiments in [4]. For example, given even q and square cells, $p \equiv q+2$ is often optimal. In this case, we have computed $c_2(q)$ explicitly for $q = 2$–20; a tight estimate is $c_2(q) \cong 15q^{1.4}$.

We assume that the coefficients a_{ij} and a_0 are constant. Then

$$\|\mathcal{T}_q^{ij}(D_{\boldsymbol{n}_{ij}}u)\|_{ij} \equiv \|\mathcal{T}_{n_{ij}}^{ij}(D_{\boldsymbol{n}_{ij}}u)\|_{ij} \leqslant 0.66 \times h^{q+1}(0.7(q+2))^{-(q+3/2)}\|(D_{\boldsymbol{n}_{ij}}u)^{q+1}\|_{ij},$$

where $(D_{\boldsymbol{n}_{ij}}u)^{q+1}$ represents the $(q+1)$st tangential derivative of $D_{\boldsymbol{n}_{ij}}u$ on Γ_{ij}.

For our cells, we have for $v \in H^{p+2}(\Omega)$ (and for $h \leqslant 3$ and $p \geqslant 2$),

$$\|\mathcal{Q}_p(v)\|_H \leqslant h^p(0.5p)^{-p}[|v|_{p+1} + |v|_{p+2}], \tag{7}$$

where, for example,

$$|v|_{p+1}^2 = \sum_{|\alpha|=p+1} \|D^\alpha v\|_0^2.$$

If we use these estimates in Theorem 1.3., with q and p replacing $[n]$ and $[m]$, and $n_f \leqslant 4$, we get

$$\begin{aligned} c_4\|u - u_{q,p}\|_H \leqslant{}& (M + M_1)c_2(q)h^{p-1}(0.5p)^{-p}[|u|_{p+1} + |u|_{p+2}] \\ &+ 2.64c_1\sqrt{N}h^{q+1/2}(0.7(q+2))^{-(q+3/2)} \max \|(D_{\boldsymbol{n}_{ij}}u)^{q+1}\|_{ij} \\ &+ c_1h^{-1/2}M_1\left\{\sum_{\Gamma_{ij}} \|\gamma_{ij}(u_{q,p}) - \gamma_{ji}(u_{q,p})\|_{ij}^2 + \sum_{\Gamma_{i0}} \|\gamma_{i0}(u_{q,p})\|_{i0}^2\right\}^{1/2}. \end{aligned} \tag{8}$$

If we are subdividing the unit square into cells of side h, the number of cells $N \cong 1/h^2$. Due to the resulting decrease in the size of Γ_{ij}, we might expect $\|(D_{\boldsymbol{n}_{ij}}u)^{q+1}\|_{ij}^2$ to decrease by a factor h. Then the h-dependency of the second error estimate term containing the normal derivative of the solution on the interfaces would be $h^{-1}h^{q+1/2}h^{1/2} = h^q$. This estimate can be made rigorous by the methods used in [4].

We report on two numerical tests of the theory.

Example 1. We approximate the solution to the following convection–diffusion problem from [14]:

$$-\sigma\Delta u + \boldsymbol{a}\cdot\nabla u = 0,$$

where $\boldsymbol{a} = (a_1, a_2)$ is a constant unit vector giving the direction of the flow. The domain is an isosceles trapezoid with base on interval $[0, 1]$ on the x-axis, acute angles of size $\pi/4$ and top on the line $y = 0.25$. Vector $\boldsymbol{a}$ points upward and makes an angle of 75° with the x-axis, so that the flow enters the domain across the x-axis. There is considerable interest in convection-dominated flows [12, 14]; we set σ to be 0.0125 in this first example. The boundary data is defined by the intended solution $u(x, y) \equiv v(s(x, y), t(x, y))$, where $s = a_1x + a_2y$, $t = a_2x - a_1y$, and

$$v(s, t) = e^{-\lambda s}\sin(\omega t)$$

with $\omega = 2\pi/a_2$ and $\lambda = (\sqrt{1 + 4\omega^2\sigma^2} - 1)/(2\sigma)$.

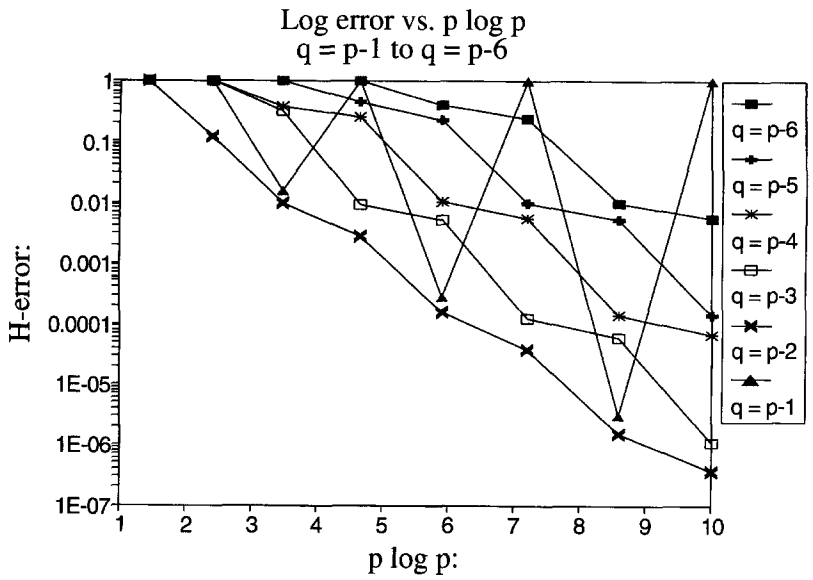

Fig. 1. Log of the H-norm of errors vs. $p \log p$ for $q = p - i$.

Our first estimate is made on a single triangular cell containing the trapezoid; we subsequently partition the trapezoid into 4 and then 12 triangles, where the lengths of the base(s) of the triangle(s) are 1, 0.5 and 0.25. We study the effect of varying h, q and p on the accuracy of the approximation.

Fig. 1 plots the H-norm of the errors for p and $q = p - i$, $p = 3, \ldots, 10$ and i ranging from 1 to 6 when the domain is partitioned into three triangles.

The approximation fails for even p when $q = p - 1$ since here we have used more than enough collocations to force continuity; there is a dependency among the rows of $\boldsymbol{M}$. The optimal value for q is $p - 2$ (when we use a partition into triangles); thus, requiring that the approximation be continuous does not seem to give any advantage. In all cases we computed the interface error

$$\left\{ \sum_{\Gamma_{ij}} \|\gamma_{ij}(u_{n,m}) - \gamma_{ji}(u_{n,m})\|_{ij}^2 + \sum_{\Gamma_{i0}} \|\gamma_{i0}(u_{n,m})\|_{i0}^2 \right\}^{1/2},$$

which is part of the general error estimate; in our tests the interface error is less than $\frac{1}{20}$ of the computed error. For example, when $p = 10$ and $q = 8$, the computed error is 0.36×10^{-6} while the interface error is only 0.64×10^{-8}. The regression equation for the error when $q = p - 2$ is

$$H\text{-error} \cong 0.94(0.46p)^{-p}.$$

This agrees well with the estimate in (8).

The h-dependency for the error shown is Fig. 2, where q is set to $p - 2$. The approximate slopes of the lines are shown on the right; they are about $q + \frac{1}{2}$; they are the powers of h in our regression

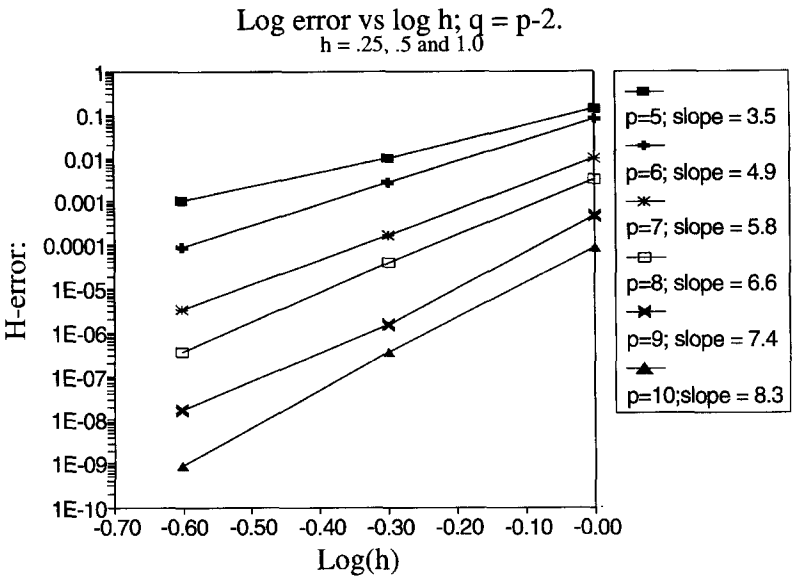

Fig. 2. Log error vs. $\log h$; $q = p - 2$; $h = 0.25$ 0.5 and 1.

model for the error of form H-error $\cong Ch^s$; thus, our experiments give error of form $Ch^{q+0.5}$, which is consistent with our theory (8).

Example 2. Traditional Galerkin methods can encounter difficulty when there is a boundary layer that can form with such convection-dominated flows [13]. We study such a problem here. Our equation is

$$-\sigma\Delta u + u_y = -\left(\frac{1}{2} + 2\sigma\pi^2(y + 1 + C_1)\right)\sin(2\pi x)$$

with domain $(0,1)\times(0,1)$ and $C_1 \equiv 2[\exp(1/\sigma)\sinh((1+16\sigma^2\pi^2)^{1/2}/\sigma) - 1]^{-1} \cong 4\exp(-2/\sigma)$, negligible for small σ. Dirichlet boundary data is zero on $y = 1$ and $x = 0$ and $x = 1$; on $y = 0$, the inflow portion of Γ, the boundary values are essentially $-0.5\sin(2\pi x)$ (with small sigma); more precisely, they are given by the intended solution

$$u(x,y) = \left[\frac{\exp((y+1)/2\sigma)\sinh(\sqrt{1+16\sigma^2\pi^2}(y+1)/2\sigma) - 1}{\exp(1/\sigma)\sinh(\sqrt{1+16\sigma^2\pi^2}/\sigma) - 1} - \frac{y+1}{2}\right]\sin(2\pi x),$$

evaluated at $y = 0$. There is a boundary layer at $y = 1$ of width $O(\sigma\ln(2/\sigma))$.

Fig. 3 shows two approximations when $\sigma = 0.005$, with $p = 10$ and $q = 7$. In Fig. 3(a) we partition the unit square into 4 congruent rectangles with base 0.25 and height 1. The boundary layer causes considerable distortion in the approximation, causing the un-physical oscillation typical of such problems [12, 14]. The width of the boundary layer is about $0.03 \cong 0.005\ln(2/0.005)$. In Fig. 3(b) we show the approximation when we use 8 cells, where each of the 4 cells above is partitioned into two

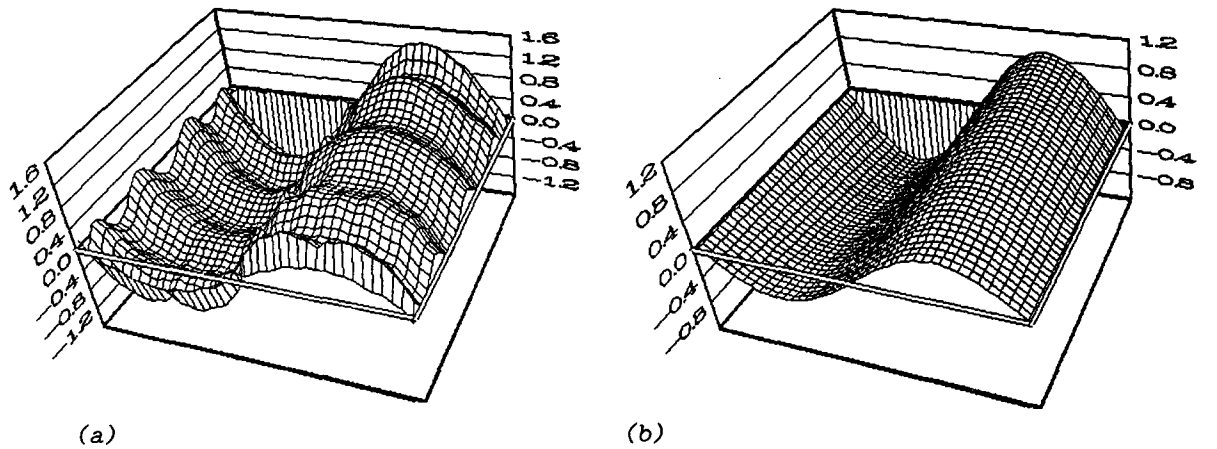

Fig. 3. Approximations to the boundary layer problem.

rectangles; a large one of width 0.25 and height 0.97 and a thin one of height 0.03 to accommodate the boundary layer. The result in Fig. 3(b) has maximum error less than 0.001.

The choice of optimal meshes for a similar problem with boundary layers when using the *hp* finite element method is treated in [16]; these results suggest directions for further investigations of the methods we have discussed in this paper.

References

[1] E. Anderson et al., LAPACK Users' Guide, SIAM, Philadelphia, PA, 1991.
[2] I. Babuška, M.R. Dorr, Error estimates for the combined *h* and *p* versions of the finite element method, Numer. Math. 37 (2) (1981) 257–277.
[3] I. Babuška, M. Suri, The *p* and *h–p* versions of the finite element method, an overview, Comput. Meth. Appl. Mech. Eng. 80 (1–3) (1990) 5–26.
[4] M. Cayco, L. Foster, H. Swann, On the convergence rate of the cell discretization algorithm for solving elliptic problems, Math. Comp. 64 (212) (1995) 1397–1419.
[5] J. Dongarra et al., LINPACK Users Guide, SIAM, Philadelphia, PA, 1979.
[6] M.R. Dorr, On the discretization of interdomain coupling in elliptic boundary-value problems, in: T.F. Chan, R. Glowinski, J. Periaux, O.B. Widlund (Eds.), Domain Decomposition Methods, SIAM, Philadelphia, PA, 1989.
[7] J. Greenstadt, Cell discretization, in: J.H. Morris (Ed.), Conference on Applications of Numerical Analysis, Lecture Notes in Mathematics, 228, Springer, New York, 1971, pp. 70–82.
[8] J. Greenstadt, The cell discretization algorithm for elliptic partial differential equations, SIAM J. Sci. Statist. Comput. 3 (3) (1982) 261–288.
[9] J. Greenstadt, The application of cell discretization to nuclear reactor problems, Nucl. Sci. Eng. 82 (1982) 78–95.
[10] J. Greenstadt, Cell discretization of nonselfadjoint linear elliptic partial differential equations, SIAM J. Sci. Stat. Comput. 12 (5) (1991) 1074–1108.
[11] P. Grisvard, Elliptic problems in non-smooth domains, Pitman, UK, 1985.
[12] B. Guo, I. Babuška, The *h–p* version of the finite element method. Part 1: The basic approximation results. Part 2: General results and applications, Comput. Mech. 1 21–41 (1986) 203–226.
[13] W.B. Liu, J. Shen, A new efficient spectral Galerkin method for singular perturbation problems, J. Sci. Comput. 11 (1996) 411–437.

[14] P.A. Raviart, J.M. Thomas, Primal hybrid finite element methods for second order elliptic equations, Math. Comp. 31 (138) (1977) 391–413.
[15] G.R. Richter, An explicit finite element method for convection-dominated steady state convection–diffusion equations, SIAM J. Numer. Anal. 28 (3) (1991) 744–759.
[16] C. Schwab, M. Suri, The p and hp versions of the finite element method for problems with boundary layers, Math. Comp. 1997, to appear.
[17] H. Swann, On the use of Lagrange multipliers in domain decomposition for solving elliptic problems, Math. Comp. 60 (201) (1993) 49–78.
[18] H. Swann, Error estimates using the cell discretization method for some parabolic problems, J. Comput. Appl. Math. 66 (1996) 479–514.
[19] J. Wloka, Partial Differential Equations, Cambridge University Press, Cambridge, 1987.